1984

The Composite Bound Method (CBM) for Computing Throughput Bounds in Multiple Class Environments

Teemu Kerola

Report Number:
84-475

The Composite Bound Method (CBM) for Computing Throughput Bounds in Multiple Class Environments

Teemu Kerola

Department of Computer Sciences

Purdue University

West Lafayette, Indiana 47907

CSD-TR 475

March 13, 1984

Revised August 27, 1984

ABSTRACT

This paper introduces a new method for calculating upper bounds on system throughput rates in multiple class closed queueing network models when any lower bounds are known. The lower (pessimistic) bounds used here are multiple class balanced job bounds. The new technique for calculating upper (optimistic) bounds is independent of the system load. The space-time computational costs are linear in the product of the number of classes and devices in the model.

Keywords: throughput bounds, multiple class queueing network models, asymptotic analysis, product form networks

1. Introduction

System throughput rate, the rate at which a system completes units of computational work, is often useful for assessing overall system performance. Queueing network models

This work was supported in part by an Academy of Finland grant (10179/711/80) and a David Ross grant from Purdue University, West Lafayette, Indiana.

have usually been used to calculate these throughput rates. Unfortunately, as the number of devices and/or the number of job classes increases, the cost of calculating throughput rates increases, to the point of rendering this approach impractical in some situations. An alternative to this so-called exact solution of models is to devise less expensive techniques which yield estimates or bounds for the throughput rates. Examples of this alternative approach include asymptotic bounds analysis (ABA) [Muntz & Wong 1974] and balanced job bounds (BJB) [Zahorjan et al 1982]. While these techniques satisfy the needs expressed above, in many cases the bounds, particularly the upper bounds, are so far away from the actual rates that the usefulness of these bounds is diminished.

This paper presents a new method for computing upper bounds for throughput rates for specified job classes, given some lower bounds for these rates exist. Not only are the computational costs much less than the cost of computing the actual rates, but the accuracy (spread between the lower and upper bounds) is usually better (tighter) than other methods for producing throughput bounds. The setting for this analysis is closed queueing network models of computing systems with multiple job classes.

The paper opens by introducing the notation, assumptions, and laws which serve as the basis of the new technique. Section 2 develops the estimates for lower bounds on the throughput rates; it also presents the new Composite Bound Method (CBM) for calculating composite throughput upper bounds (CUB's). An algorithm for computing CUB's is given in Section 3. The properties of composite bounds are discussed in Section 4. Section 5 contains examples, and the summary of results is in Section 6. Appendix A contains a list of all used acronyms and Appendix B gives the proof for the Composite Bound Theorem.

Consider a closed queuing network of K devices. The customer (job) population in the network is $\mathbf{N}=(N_1, N_2, ..., N_R)$, where R is the number of customer classes, and N_r is the number of jobs in class r. The total population is $N = N_1 + N_2 + ... + N_R$. Each customer in

class r requires L_{kr} amount of work at device k before it departs from the system. The value L_{kr} is called the *loading* for device k in class r. In some other references (e.g. [Denning & Buzen 1978]) the term L_{kr} is denoted by $V_{kr}S_{kr}$, where the *visit ratio* V_{kr} is the mean number of times each class r job visits device k, and the *service time* S_{kr} is the mean service requirement of a class r job for each visit at device k. All the summations, maximums, and minimums used are over all devices (index k) or all classes (indexes r and s) unless marked otherwise.

For class r the *maximum loading* is

$$L_{br} = \max_k L_{kr}, \tag{1}$$

the *average loading* is

$$L_{ar} = \frac{\sum_k L_{kr}}{K}, \tag{2}$$

and the *minimum loading* is

$$L_{mr} = \min_k L_{kr}.$$

The *minimum response time* for class r is

$$R_{0r} = \sum_k L_{kr}. \tag{3}$$

The *system throughput* $X_{0r}(\mathbf{N})$ is the rate at which class r customers pass through the system. The *device throughput* $X_{kr}(\mathbf{N})$ is the rate at which class r customers pass through the device k. In the single class case the corresponding measures are denoted with V_k, S_k, L_b, L_a, L_m, R_0, $X_0(\mathbf{N})$, and $X_k(\mathbf{N})$.

In the following the parameter **N** may be omitted from $X_{0r}(\mathbf{N})$ or from any other function of **N** if its value is obvious. All symbols for upper (lower) bounds in this presentation have a superscript ending with + (−). It is assumed that all service times S_{kr} are load independent.

Little's Law [Little 1961] states that the number of customers in any system equals the system throughput times the system response time. This formula can be applied to the network as a whole or to any single device in the network. It applies also to any single class in a multiple class system. The Forced Flow Law [Denning & Buzen 1978] states that the throughput for class r through device k equals the system throughput for class r times the visit ratio:

$$X_{kr}(\mathbf{N}) = X_{0r}(\mathbf{N})V_{kr}. \tag{4}$$

The Utilization Law [Denning & Buzen 1978] states that the class r utilization $U_{kr}(\mathbf{N})$ at device k equals class r throughput at device k times class r service time at device k:

$$U_{kr}(\mathbf{N}) = X_{kr}(\mathbf{N})S_{kr}. \tag{5}$$

No device can have utilization more than 1.0. From (4) and (5) one obtains

$$U_{kr}(\mathbf{N}) \overset{(5)}{=} X_{kr}(\mathbf{N})S_{kr} \overset{(4)}{=} X_{0r}(\mathbf{N})V_{kr}S_{kr} = X_{0r}(\mathbf{N})L_{kr}. \tag{6}$$

2. System Throughput

If the network has a product form solution [Basket et al 1975], we can obtain the exact system throughput for every class, but this may require large amounts of computation even for a moderate number of job classes and devices. Both the convolution method [Reiser & Kobayashi 1975] and the mean value analysis (MVA) [Reiser & Lavenberg 1980] for solving multiple class networks have time complexity [Lazowska et al 1984]

$$O\left(RK\prod_{r=1}^{R}(N_r+1)\right)$$

and space complexity

$$O\left(K\prod_{r=2}^{R}(N_r+1)\right).$$

Because computing the exact throughput may be too expensive or practically impossible, one would like to be able to obtain relatively tight throughput bounds and other performance

measures obtainable from the class throughputs, with significantly less computational work.

The asymptotic bound analysis (ABA, bottleneck analysis) produces linear optimistic asymptotic bounds for the throughput of single class networks:

$$X_0^{ABA+}(\mathbf{N}) = \min\{\frac{N}{R_0}, \frac{1}{L_b}\}.$$

The single class bottleneck device is the device b with the highest loading L_b. The corresponding multiple class bounds

$$X_{0r}^{ABA+}(\mathbf{N}) \stackrel{def}{=} \min\{\frac{N_r}{R_{0r}}, \frac{1}{L_{br}}\}. \tag{7}$$

can be used as relatively loose optimistic bounds. Intuitively this bound should not be as good because it does not take into consideration other classes. Later it is shown that (7) is the asymptotic bound for class r throughput. The time complexity to compute ABA bounds (7) for all classes is $O(KR)$.

Zahorjan *et al* [Zahorjan et al 1982] introduced nonlinear throughput bounds. Their Balanced Job Bounds are based on comparing the existing network to balanced networks where the loading at each device is either the maximum, minimum, or the average loading in the original network. The Simple Balanced Job Bounds for multiple class networks are

$$X_{0r}^{SBJB-}(\mathbf{N}) \stackrel{def}{=} \frac{N_r}{(N+K-1)L_{br}} \le X_{0r}(\mathbf{N}) \le \frac{N_r}{(N+K-1)L_{mr}} \stackrel{def}{=} X_{0r}^{SBJB+}(\mathbf{N}). \tag{8}$$

The complexity of computing (8) for all classes is $O(KR)$. They also presented tighter bounds for single class networks:

$$\frac{N}{R_0+(N-1)L_b} \le X_0(N) \le \frac{N}{R_0+(N-1)L_a}. \tag{9}$$

The lower bound of (9) can be extended to the multiple class balanced job bounds (BJB's):

$$X_{0r}^{BJB-}(\mathbf{N}) \stackrel{def}{=} \frac{N_r}{R_{0r}+(N-1)L_{br}} \le X_{0r}(\mathbf{N}). \tag{10}$$

This is proven easily in a similar fashion as (9) was proven in [Zahorjan et al 1982].

Notice that

$$X_{0r}^{SBJB-} \overset{(8)}{=} \frac{N_r}{(N+K-1)L_{br}} = \frac{N_r}{KL_{br}+(N-1)L_{br}}$$

$$\overset{(1)}{\leq} \frac{N_r}{KL_{ar}+(N-1)L_{br}} \overset{(2)}{=} \frac{N_r}{R_{0r}+(N-1)L_{br}} \overset{(10)}{=} X_{0r}^{BJB-} \tag{11}$$

i.e. the bound X_{0r}^{BJB-} from (10) is always closer to $X_{0r}(\mathbf{N})$ than the bound X_{0r}^{SBJB-} from (8).

The multiple class extension of the upper bound in (9),

$$\frac{N_r}{R_{0r}+(N-1)L_{ar}}, \tag{12}$$

is not a valid upper bound. Later an example is shown where the value using (12) is less than the actual throughput for class r.

Kriz [Kriz 1984] proved (10) in extended form for networks with delay terminals and derived the corresponding upper bound using L_{mr} instead of L_{br}. His upper bound crosses $1/L_{br}$ and so the minimum of these two bounds is used. Although the bound introduced in this paper is always less than $1/L_{br}$, it can be larger than Kriz's upper bound for small N_r values as is shown in Example 3. Kriz's lower bound reduces to the BJB lower bound for the networks considered in this paper.

Our main contribution is the Composite Bound Theorem which gives a new composite upper bound $X_{0r}^{CUB+}(\mathbf{N})$ for the system throughput in any given class:

$$X_{0r}(\mathbf{N}) \leq X_{0r}^{CUB+}(\mathbf{N}) \overset{def}{=} \min_k \frac{1-\sum_{s\neq r} X_{0s}^{BJB-}(\mathbf{N})\, L_{ks}}{L_{kr}}. \tag{13}$$

It is proven in Appendix B. The composite upper bound method can be used with multiple class networks to produce non-linear upper bounds for the throughput for class r assuming that lower bounds for the throughputs for other classes are known. In this paper the lower throughput bounds are computed with (10), but (13) applies to any other lower bound as well. If one uses zero as the lower bound instead of (10), (13) reduces to

$$X_{0r}(\mathbf{N}) \leq \min_k \frac{1}{L_{kr}} = \frac{1}{L_{br}},$$

where the upper bound is just the single class bottleneck analysis asymptotic bound from (7). The following section gives the reasoning behind (13) in an algorithmic form.

3. The Composite Bound Method (CBM)

Let $A \in \{1,..., R\}$ denote any class. We want to find an upper bound for the throughput for class A given a joint multiprogramming load **N**.

The basic idea is as follows: Given the minimum throughput bounds for all other classes, compute the minimum utilization that the jobs in other classes will contribute at each device. The unused utilization at each device is then the maximum utilization that class A jobs can obtain at that device. These maximum device utilizations give us maximum throughputs for each device and for the system. The *CBM bottleneck device* for class A (hereafter bottleneck device for class A) is the device that gives the lowest system throughput bound, and the *CUB* is the corresponding system throughput bound. Note that the bottleneck device is defined separately for each joint multiprogramming load **N**. In single class bottleneck analysis, the bottleneck device is determined solely by the network loadings L_{kr} and each device has maximum utilization 1.0. The algorithm to compute the composite upper bound for class A system throughput is presented in Figure 1.

Notice that each *Composite Device Upper Bound* $X_{0A}^{CUBk+}(\mathbf{N})$, as defined in step A4 of the CUB algorithm (Figure 1), gives an upper bound for the class A system throughput, and that in step A5 the lowest of them is selected as the composite upper bound. The method is similar to the single class bottleneck analysis, where a horizontal bound corresponding to every device is computed and the smallest of them is selected as the ABA upper throughput bound. If the bounds X_{0A}^{CUBk+} are plotted as functions of N_A (while keeping all other N_r's constant) we obtain a set of curvilinear bounds, one for each device, which can cross each other. The CUB

Input: Loadings L_{kr}, joint multiprogramming load N, and class $A \in \{1,..., R\}$

Output: Composite upper bound $X_{0A}^{CUB+}(\mathrm{N})$ for class A throughput

Algorithm CUB:

A1. Use (14) to find a lower bound $X_{0r}^{CUB-}(\mathrm{N})$ for X_{0r} for every class r.

A2. Use (6) to compute a composite lower limit U_{kA}^{other-} for device utilization at each device k due to all other classes but A:

$$U_{kA}^{other-} \stackrel{def}{=} \sum_{r \neq A} X_{0r}^{CUB-}(\mathrm{N})\, L_{kr}.$$

A3. Compute the remainder of each device capacity i.e. the maximum utilization U_{kA}^{CUB+} that class A jobs could use at each device k:

$$U_{kA}^{CUB+} \stackrel{def}{=} (1 - U_{kA}^{other-}).$$

The term U_{kA}^{CUB+} is the composite upper bound for the utilization at device k.

A4. Use (6) again to obtain upper bounds X_{0A}^{CUBk+} for class A throughput for each device k:

$$X_{0A}^{CUBk+} \stackrel{def}{=} \frac{U_{kA}^{CUB+}}{L_{kA}} \quad for\ all\ k = 1,..., K.$$

A5. Select the smallest of them as the composite upper bound for class A throughput:

$$X_{0A}^{CUB+}(\mathrm{N}) \stackrel{def}{=} \min_k X_{kA}^{CUBk+} = \min_k \frac{U_{kA}^{CUB+}}{L_{kA}} = \min_k \frac{1 - U_{kA}^{other}}{L_{kA}} = \min_k \frac{1 - \sum_{r \neq A} X_{0r}^{CUB-}(\mathrm{N})\, L_{kr}}{L_{kA}}.$$

Figure 1: The Composite Upper Bound Algorithm

at any given point N_A is the smallest device bound at that point.

The method described above can easily be extended to compute the CUB's for all classes. The space-time computational complexity (to compute the CUB's for all classes) is still $O(KR)$ when the BJB lower bounds (10) are used.

4. Properties

In this section some properties of the composite bounds are discussed. First, note that

$$X_{0r}^{CUB+}(\mathbf{N}) \overset{(13)}{=} \min_k \frac{1-\sum_{s\neq r} X_{0s}^{BJB-}(\mathbf{N})\, L_{ks}}{L_{kr}} \leq \min_k \frac{1}{L_{kr}} = \frac{1}{L_{br}} \tag{15}$$

$$\overset{(7)}{=} X_{0r}^{ABA+}(\mathbf{N}), \qquad \text{if } N_r \geq N_r^* \overset{def}{=} \frac{R_{0r}}{L_{br}}.$$

The point N_r^* is the single class *saturation point* [Denning & Buzen 1978], where $N_r/R_{0r} = 1/L_{br}$.

Now, (11), (13), and (15) lead to

$$X_{0r}^{SBJB-}(\mathbf{N}) \overset{(11)}{\leq} X_{0r}^{BJB-}(\mathbf{N}) \overset{(13)}{\leq} X_{0r}(\mathbf{N}) \overset{(13)}{\leq} X_{0r}^{CUB+}(\mathbf{N})$$

$$\overset{(15)}{\leq} X_{0r}^{ABA+}(\mathbf{N}), \quad \text{if } N_r \geq N_r^*. \tag{16}$$

The outermost functions in (16) converge to the same value:

$$\lim_{N_r\to\infty} X_{0r}^{SBJB-}(\mathbf{N}) \overset{(8)}{=} \lim_{N_r\to\infty} \frac{N_r}{(N_1+..+N_r+..+N_R+K-1)L_{br}} = \frac{1}{L_{br}} \overset{(7)}{=} \lim_{N_r\to\infty} X_{0r}^{ABA+}(\mathbf{N}). \tag{17}$$

From (16) and (17), we obtain

$$\lim_{N_r\to\infty} X_{0r}^{BJB-}(\mathbf{N}) = \lim_{N_r\to\infty} X_{0r}(\mathbf{N}) = \lim_{N_r\to\infty} X_{0r}^{CUB+}(\mathbf{N}) = \frac{1}{L_{br}}. \tag{18}$$

Both the BJB lower bound $X_{0r}^{BJB-}(\mathbf{N})$ and the corresponding composite upper bound $X_{0r}^{CUB+}(\mathbf{N})$ converge to the same value $1/L_{br}$ when N_r increases and other class populations remain constant. In a multiple class system the asymptotic throughput value for any class system throughput is the same as if the jobs of that class were alone in the system. This corresponds to the intuitive notion that if the number of jobs in one class is increased while keeping the multiprogramming level in other classes the same the effect of jobs in other classes decreases and eventually vanishes.

With simple analysis we can transform the CUB to

$$X_{0r}^{CUB+}(\mathrm{N}) = X_{0r}^{BJB-} + \min_k \frac{1 - \sum_s X_{0s}^{BJB-}(\mathrm{N}) L_{ks}}{L_{kr}}, \tag{19}$$

where $1 - \sum_s X_{0s}^{BJB-}(\mathrm{N}) L_{ks}$ is the unknown utilization capacity at device k and

$$\min_k \frac{1 - \sum_s X_{0s}^{BJB-}(\mathrm{N}) L_{ks}}{L_{kr}} \tag{19}$$

is the gap between the lower throughput bound and the corresponding CUB. The gap becomes smaller when loadings L_{kr} in the class of interest become larger or the known lower bounds used (here X_{0r}^{BJB-}) become larger i.e. more accurate. The bounds X_{0r}^{BJB-} work best with balanced workloads, and the more unbalanced the class is, the more greater the deviation of the factor L_{br} in (10), and the more pessimistic the lower bounds X_{0r}^{BJB-} are.

Even in the case that the lower throughput bounds have no error the gap (19) can be large. If the network is under-utilized, i.e. no device is fully utilized, then all that unused utilization is always fully assigned to the class of interest when computing the CUB. The gap becomes smaller when device utilizations become larger.

Because the CUB for the class throughput is based on the throughputs of other classes, at least two classes are required for the CUB to be effective. For a single class network the CBM reduces to the asymptotic bounds analysis.

The CUB can be obtained for *any* lower bounds which can be computed for the class throughputs. Thus, if a new and better lower bound (better than X_{0r}^{BJB-}) can be found, a corresponding new and tighter CUB upper bound is also available. Given any method to compute the lower bounds, the CUB upper bounds are applicable whenever the lower bounds exist.

The CUB's can be computed directly for any load. There is no need to compute the bounds for any smaller loads before obtaining the values for larger loads. The time and space complexities depend only on the number of queues and the number of classes but not on the

class populations.

The ABA upper bound N_r/R_{0r} for the class r throughput can be better than the CUB upper bound for small N_r values. We can consider the point N_r^{CB*} where the bound N_r/R_{0r} crosses the CUB, to be the *Composite Bound Saturation Point* for class r.

Kriz's upper bound [Kriz 1984], which is always at least as good (low) as the ABA bound, can also be better than the CUB, but again only up to a point. Kriz's upper bound reaches $1/L_{br}$ at

$$N_r^{Kr*} \stackrel{def}{=} \frac{R_{0r} + (N_r^{other} - 1)L_{mr}}{(L_{br} - L_{mr})},$$

where N_r^{other} is the total population in all other classes but r. So, at least for all $N_r \geq N_r^{Kr*}$ the CUB is guaranteed to be a better bound. It is advisable, especially for small network populations, to compute both bounds and use the smallest.

5. Examples

Examples 1 and 2 illustrate the CUB algorithm. Example 3 compares the CUB to Kriz's upper bound.

Example 1

Consider a simple two-class two-device network with loadings

$$L_{11} = 0.9 \quad L_{12} = 0.9$$

$$L_{21} = 0.1 \quad L_{22} = 1.0.$$

We are interested in the class 1 throughput when the workload is $\mathbf{N}=(N_1, N_2)=(2, 3)$. First, from the above loadings we compute the values

$$R_{01} = 1.0 \quad R_{02} = 1.9$$
$$L_{b1} = 0.9 \quad L_{b2} = 1.0.$$

The CUB algorithm is now applied to obtain the class 1 throughput bounds:

1) First, compute the CUB lower bounds for the class throughputs:

$$X_{01}^{BJB-}(\mathrm{N}) = \frac{N_1}{R_{01}+(N-1)L_{b1}} = \frac{2}{1.0+4*0.9} = 0.435 \leq X_{01}(\mathrm{N})$$

$$X_{02}^{BJB-}(\mathrm{N}) = \frac{N_2}{R_{02}+(N-1)L_{b2}} = \frac{3}{1.9+4*1.0} = 0.508. \leq X_{02}(\mathrm{N})$$

The term $X_{01}^{BJB-}(\mathrm{N})$ is a lower bound for $X_{01}(\mathrm{N})$ and $X_{02}^{BJB-}(\mathrm{N})$ is used to compute an upper bound for $X_{01}(\mathrm{N})$.

2) Next, compute the minimum utilization that class 2 jobs cause at both devices:

$$U_{12}^{other-} = X_{02}^{BJB-} L_{12} = 0.508*0.9 = 0.457$$

$$U_{22}^{other-} = X_{02}^{BJB-} L_{22} = 0.508*1.0 = 0.508.$$

3) Now, obtain the remaining utilization capacity at both devices:

$$U_{11}^{CUB+} = (1-U_1^{other-}) = 0.543$$

$$U_{21}^{CUB+} = (1-U_2^{other-}) = 0.492.$$

4-5) Last, find the device that would first reach its maximum utilization, and the corresponding maximum class 1 system throughput. So, with multiprogramming load N=(2, 3), the maximum throughput is

$$X_{01}^{CUB+}(\mathrm{N}) = \min_k \frac{U_{k1}^{CUB+}}{L_{k1}} = \min\left(\frac{0.543}{0.9}, \frac{0.492}{0.1}\right) = \min(0.603, 4.92) = 0.603,$$

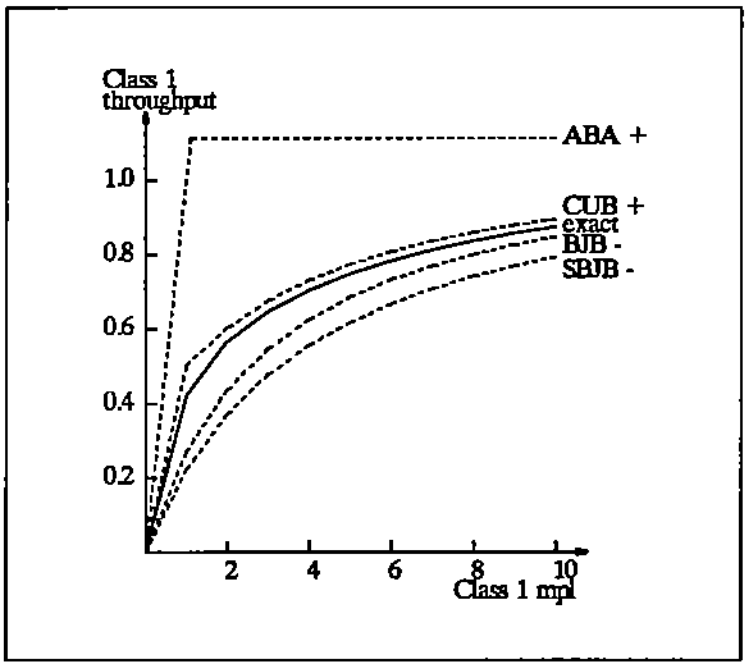
and the class 1 bottleneck device is the device 1.

The exact MVA solution is X_{01}=0.565, and the ABA upper bound from (7) is 1.11.

This same example can be used as an counter example to show that the upper bound of formula (9) can not be expanded to the multiple class case. The exact class 2 throughput is 0.538 whereas the value from (12) is 0.526.

Example 2

This example concerns the same network as Example 1. The class 1 throughput bounds and the exact values were computed for N_1=1,..., 15 while keeping N_2 at constant 3. The resulting bounds are shown with dotted lines and the exact value with a solid line in Graph 1.



Graph 1: The Example 1 Network: Throughput X_{01} vs N_1

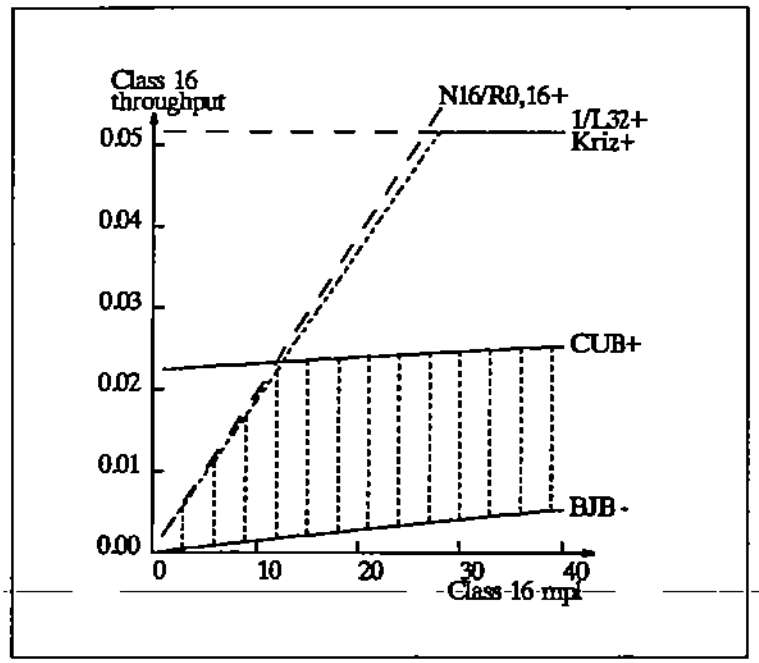
Class 2 population $N_2 = 3$	
ABA+ (dashed line)	ABA upper bound from (7)
CUB+ (dashed line)	CUB upper bound from (13)
exact (solid line)	Exact solution from multiple class MVA model
BJB- (dashed line)	BJB lower bound from (10)
SBJB- (dashed line)	SBJB lower bound from (8)

Example 3

A number of random networks were generated while evaluating the effectiveness of the CUB's. We chose the number devices and classes in the network and selected randomly (from uniform distributions) the class r of interest, the work load at every device for each class, the relative speeds of every device, and the class populations in all other classes but r. In general, Kriz's upper bound seems to be better (lower) for small N_r's, but reaches soon its maximum limit $1/L_{br}$ whereas the CUB grows slowly and remains well below $1/L_{br}$ all the time.

This network is one such random network with parameter values K=100, R=20, r=16,

$N_s \in Uniform[1,40]$ for $s \neq 16$, $WorkL_{ks} \in Uniform[0.001,1.0]$ for all k and all s, $N_s \in Uniform[1,40]$ for all $s \neq 16$, and $Speed_k \in Uniform[1,20]$ for all k, where $WorkL_{ks}$ is the work load at device k for class s jobs and $Speed_k$ is the relative speed of device k. The loadings were set to $L_{ks} = WorkL_{ks} Speed_k$. The population in class $r = 16$ was varied from 1 to 5000, but the bounds were plotted only for populations up to 40. The bounds are shown in Graph 2.



Graph 2: The Example 3 Network: Throughput $X_{16}(N)$ vs N_{16}

N16/R0,16+ (dashed line)	non-horizontal ABA bound in (7)
1/L32+ (dashed line)	horizontal ABA bound in (7)
Kriz+ (dashed line)	Kriz's upper bound (see Section 2)
CUB+ (solid line)	CUB upper bound from (13)
BJB- (solid line)	BJB lower bound from (10)

In this case the CUB becomes a better upper bound than Kriz's or the ABA upper bound at N_{16}=13. Kriz's upper bound reaches its maximum $1/L_{32,16}$=0.0516 at N_{16}=28. The CUB is initially (N_{16}=1) relatively high and grows slowly. At point N_{16}=5000 (not shown in Graph 2) the CUB $X_{0,16}^{CUB+}$=0.0497 is still 4% below its asymptotic value 0.0516.

The accuracy of the CUB depends how close the lower bounds (here BJBs) are to the actual throughputs and how heavily loaded the system is. We can obtain some information of the combined effect of these two factors by computing the known minimum utilization of the bottleneck device (here device 32 for all $N_{16} \in [1..5000]$) for the class of interest (here class 16):

$$U_{32}^{MIN^-}(\mathbf{N}) = \sum_{s} X_{0s}^{BJB^-} L_{32,s}.$$

In this case $U_{32}^{MIN^-}(\mathbf{N})$ is 60-62% for $N_{16} \in [1..40]$. We do not know how much of the remaining unassigned utilization (38-40%) is unused or how much of it is attributed to the differences between the true throughputs and the lower bounds $X_{0r}^{BJB^-}(\mathbf{N})$. The CBM assigns all the remaining utilization for class 16 and this will result to relatively high CUBs and relatively large range for possible class 16 system throughput values. As the class 16 population is increased, $U_{32}^{MIN^-}(\mathbf{N})$ increases gradually and reaches 97% for N_{16}=5000.

6. Summary

We have developed a new method for obtaining class throughput upper bounds from class throughput lower bounds for multiple class closed queueng networks. The upper bound converges asymptotically to the corresponding single class bottleneck bound and in most cases is better than any previous bound. The composite bounds are simple to compute; in a multiple class network we can not hope for better than $O(KR)$ computational complexity if we want to incorporate all the devices and classes into the bound analysis.

The CBM bottleneck device (for class r) was defined in a multiple class system to be the device that under the current multiprogramming load gives the lowest system throughput upper bound for class r jobs. The bottleneck device in a multiple class system is not only a function of the network loadings (as is the case with single class networks) but is also a function of the current multiprogramming load. Even if we keep the population of other classes constant, the bottleneck device for the given class can change with a change in the population

of that class.

Using the composite bounds it was shown that the asymptotic single class system throughput behavior in a multiple class system is the same as if all the jobs in other classes were removed. This corresponds to the intuitive notion that if the number of jobs in one class is increased while keeping the multprogramming level in other classes the same, the effects of jobs in other classes will decrease and eventually vanish.

The method of computing optimistic class throughput bounds given pessimistic class throughput bounds can be extended to any better lower bound that would be developed in the future. In this way we have not only found a new formula to compute class throughput bounds but we have also developed a new methodology for throughput bound analysis.

A more detailed description of the CBM, its accuracy, and its applications can be found in [Kerola 1984]. Our further work includes extending the CBM to wider range of networks, including those with delay and load-dependent servers. For networks with delay servers we will apply the CBM using Kriz's lower bounds. If the Performance Bound Hierarchies [Eager & Sevcik 1983], the Throughput Bound Hierarchies [Dowdy, Perez-Davila & Stephens 1983], or Generalized Quick Bounds [Suri 1983] are extended to multiple classes we would like to explore the advantages of combining them with the CBM.

Acknowledgements. Professor Herb Schwetman helped me make the presentation more succinct and precise. Cristina Ruggieri and André B. Bondi were helpful in locating mistakes in the original manuscript. Anonymous referees helped to clarify the presentation.

Appendix A: List of used acronyms

Acronym	Meaning	Reference
ABA	Asymptotic Bound Analysis, Bottleneck Analysis	(7)
BJB	Balanced Job Bounds	(10)
CBM	Composite Bound Method	Section 3
CUB	Composite (Throughput) Upper Bound	(13)
MVA	Mean Value Analysis	Section 2
SBJB	Simple Balanced Job Bounds	(8)

Appendix B: Proof of the Composite Bound Theorem

THEOREM: Consider a closed multiple class (R classes) queueing network, which is composed of K devicess with loading L_{kr} for each class $r \in \{1,..., R\}$ at each device $k \in \{1,..., K\}$, where L_{kr} is the maximum loading in class r, and where R_{0r} is the sum of loadings in class r. Assume that the multiprogramming level in each class r is N_r; thus the joint multiprogramming load is $\mathbf{N} = (N_1, N_2, \ldots, N_R)$. Let $N = N_1 + N_2 + \ldots + N_R$ and let $r \in \{1,..., R\}$. Let X_{0s}^{LOW-} be any lower bound for class s system throughput X_{0s} which is defined for all $s \in \{1,..., R\}$. If the queueing network has a product-form solution then

$$X_{0r}(\mathbf{N}) \leq X_{0r}^{CUB+}(\mathbf{N}) \stackrel{def}{=} \min_k \frac{1 - \sum_{s \neq r} X_{0s}^{LOW-}(\mathbf{N})\, L_{ks}}{L_{kr}}.$$

PROOF: Let $k \in \{1,..., K\}$ be any device. From (6) we first deduce that the minimum utilization U_{0r}^{other-} due to all other classes but r at device k is

$$U_{kr}^{other-}(\mathbf{N}) \stackrel{def}{=} \sum_{s \neq r} X_{0s}^{LOW-}(\mathbf{N})\, L_{ks}.$$

No device can have utilization greater than 1.0. So, the maximum utilization that class r jobs can have at device k is $1 - U_{kr}^{other-}$ i.e.

$$U_{kr}(\mathbf{N}) \leq 1 - U_{kr}^{other-}(\mathbf{N}) = 1 - \sum_{s \neq r} X_{0s}^{LOW-}(\mathbf{N})\, L_{ks}.$$

Using (6) again, an upper limit is found for the maximum system throughput for class r:

$$X_{0r}(\mathbf{N}) \stackrel{(6)}{=} \frac{U_{kr}(\mathbf{N})}{L_{kr}} \leq \frac{1 - \sum_{s \neq r} X_{0s}^{LOW-}(\mathbf{N})\, L_{ks}}{L_{kr}}.$$

The preceding formula is true for all k, so

$$X_{0r}(\mathbf{N}) \leq \min_k \frac{1 - \sum_{s \neq r} X_{0s}^{LOW-}(\mathbf{N})\, L_{ks}}{L_{kr}} = X_{0r}^{CUB+}(\mathbf{N}).$$

□

References

Denning & Buzen 1978. Denning, Peter J. and Buzen, Jeffrey P., "The Operational Analysis of Queueing Network Models," *Computing Surveys* **10**(3) p. 225-261 (Sep 1978).

Dowdy, Perez-Davila & Stephens 1983. Dowdy, Lawrence W., Perez-Davila, Alfredo, and Stephens, Lindsey E., *Performance Bounds Based Upon Throughput Curve Properties (presented in Performance 83)*, North-Holland (1983).

Eager & Sevcik 1983. Eager, Derek L. and Sevcik, Kenneth C., "Performance Bound Hierarchies for Queueing Networks," *ACM Trans. on Computer Systems* 1(2) p. 99-115 (May 1983).

Lazowska et al 1984. Lazowska, Edward D., Zahorjan, John, Graham, G. Scott, and Sevcik, Kenneth C., *Computer System Analysis Using Queueing Network Models*, Prentice-Hall, New Jersey (1984).

Little 1961. Little, J.D.C., "A Proof of Queueing Formula L = λW," *Oper. Res.* 9 p. 383-387 (1961).

Muntz & Wong 1974. Muntz, R.R. and Wong, J. W., "Asymptotic Properties of Closed Queueing Network Models," *Proc. 8th Princeton Conf. Information Sciences and Systems*, p. 348-352 Princeton Univ., Princeton, N.J., (March 1974).

Reiser & Kobayashi 1975. Reiser, M. and Kobayashi, H., "Queueing Networks with Multiple Closed Chains: Theory and Computational Analysis," *IBM J. Res. Devel.* 19 p. 283-294 (May 1975).

Reiser & Lavenberg 1980. Reiser, M. and Lavenberg, S. S., "Mean-Value Analysis of Closed Multichain Queuing Networks," *J. ACM* 27(2) p. 313-322 (1980).

Suri 1983. Suri, Rajan, "Generalized Quick Bounds for Performance of Queueing Networks," TR-05-83, Harward University, Center for Research in Computing Technology (1983).

Zahorjan et al 1982. Zahorjan, John, Sevcik, Kenneth C., Eager, Derek L., and Galler, Bruce, "Balanced Job Bound Analysis of Queueing Networks," *Comm. ACM* 25(2) p. 134-141 (Feb 1982).

Basket et al 1975. Baskett, Forest, Chandy, K. Mani, Muntz, Richard R., and Palacios, Fernando G., "Open, Closed, and Mixed networks of Queues with Different Classes of Customers," *J. ACM* 22(2) p. 248-260 (1975).

Kerola 1984. Kerola, Teemu, *Ph.D. Thesis*, Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907 (To appear 1984).

Kriz 1984. Kriz, Jiri, "Throughput Bounds for Closed Queueing Networks," *Performance Evaluation* 4(1) p. 1-10 (Feb 1984).